Choosing a neighbourhood in Toronto, Canada

EJ Romero

2021May30

Background

The city of Toronto has 140 neighbourhoods that is a viable location for a certain stakeholder/s, single family, which wants to potentially start a Thai restaurant business at and live nearby. I have been tasked to evaluate these 140 Toronto neighbourhoods and provide assessment based on health and safety factors for them.

Audience

The target audience (or stakeholders) are mainly for individuals (in this case single-type families [1-4 members]) who want to start a food restaurant and move in the Toronto area. The use-case and dataset are also viable for individuals who want to start an ethnic-based food restaurant (in this use-case Thai [or Asian/Pacific Islander] cuisine).

The assessment and datasets provided can help determine the best possible neighbourhood locations and provide the stakeholder/s a set of choices based on their requirements. The assessment can also provide information what the state of restaurant businesses in a neighbourhood (e.g. provide what type of restaurants are in the neighborhood).

Neighbourhood Candidates

The candidates are chosen from records provided by the City of Toronto's Open Data Platform. All 140 neighbourhoods will be assessed based on selected and satisfactory factors. Ultimately, one neighbourhood will be chosen for further assessment using the Foursquare API to determine food venues, mainly Asian food restaurants.

Methodology

The best neighbourhood will mainly be selected based on safety and health factors. These factors include neighbourhood crime rates, fire rates, hazardous incidents, pollutants released in air, carcinogenic TEP score, and health providers. Each factor is provided by the City of Toronto and have been recorded in a 2011 census. Currently, no new datasets were available during assessment except for 2016 population rates and a 2020 crime rate for each neigbourhood. These newer datasets will be utilized as updates in main 2011 datasets.

The first step will heavily rely on collecting data from Toronto's Open Data Platform. Each dataset will be cleaned and processed using Python and Excel. Once, appropriate pre-data exploration will be done to assess the quality of the data.

The second step will comprise with further data exploration and analysis. This step will assess which datasets will be used and divided into subsets for further investigation to verify safety, health, and extra factors that will be helpful for the client to understand the provided data. This step will also include clustering all Toronto 140 neighbourhoods and will be further broken down in preparation for the next step.

The third step will mainly focus on selecting the best neighbourhood in the Toronto area. Selection includes central Toronto neighbourhoods, meaning that the selection process will only select neighbourhoods on the safety and health factors - and to include neighbourhoods in "Toronto" (Toronto-named boroughs).

The fourth step will utilize the Foursquare API to determine available restaurants that may be competitive in that neighbourhood area. In these use-case scenario, competition will be based on available Thai restaurant/s (or possibly similar ethnic cuisine). Listing other restaurants in the area may still be useful data to the stakeholder to assess other plans in the neighbourhood area.

Data

The data collected for this project was provided by the City of Toronto's Open Data Platform. Resources to the data will be provided in the final notebook/presentation/report. Collected data are as follows:

Demographics
Economics
Environment
Health
Housing
Crime Rates
Safety
Total Population

Respectively, each dataset were collected on 2011 through 2020 - the majority of the data collected being on 2011. These datasets are processed (or quantified) by neighbourhood. The Demographics dataset include population, age range, and language used. The Economics dataset include home prices and total businesses. The Environment dataset include pollutant scores and tree cover (trees available/saturation). The Health dataset include DineSafe inspections, health providers, and student nutrition scores. The Housing dataset also include home pricing. Both Crime Rates and Safety datasets are divided by some of the most common infractions - assault, auto theft, breaking and entering, robbery, theft, and homicide. The Total Population dataset include total population by year and population change rate.

Please note that the neighbourhood datasets obtained from the City of Toronto reflects 140 neighbourhoods in the Toronto area. Neighbourhood names have been updated as well as each of these neighbourhoods have been assigned a "Neighbourhood Id". Some of the neighbourhoods may also be combined (or separated) with other neighbourhoods. In 2021, more neighbourhoods have been classified in Toronto - 158. The 158 neighbourhood dataset is not used in this assessment.

Data Resources

City of Toronto's Open Data Platform:
https://open.toronto.ca/

Toronto Police:
https://data.torontopolice.on.ca/

Neighbourhood Crime Rates 2011:
https://open.toronto.ca/dataset/neighbourhood-crime-rates/

Neighbourhood Crime Rates 2020:
https://data.torontopolice.on.ca/datasets/neighbourhood-crime-rates-2020-1

Toronto Environment:
https://open.toronto.ca/dataset/wellbeing-toronto-environment/

Toronto Health:
https://open.toronto.ca/dataset/wellbeing-toronto-health/

Toronto Safety:
https://open.toronto.ca/dataset/wellbeing-toronto-safety/

Toronto Housing:
https://open.toronto.ca/dataset/wellbeing-toronto-housing/

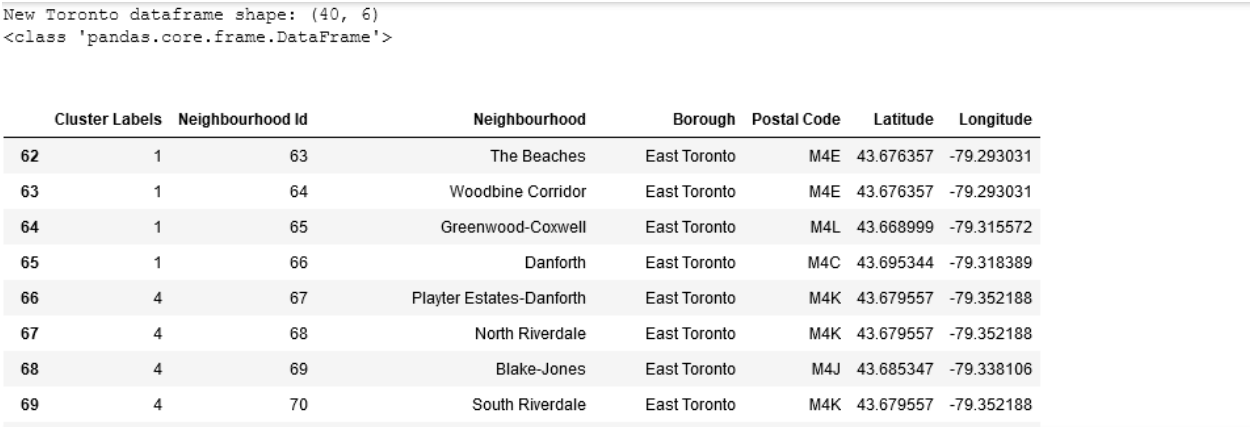
Toronto Economics:
https://open.toronto.ca/dataset/wellbeing-toronto-economics/

Toronto Demographics:
https://open.toronto.ca/dataset/wellbeing-toronto-demographics/

Canada Inflation Rate:
https://tradingeconomics.com/canada/inflation-cpi

Data Pre-processing

Data research, data cleaning, and data pre-processing took the most time in this assessment. Data processing were evaluated using both Jupyter Notebook and Microsoft Excel. The datasets provided will be utilized to assess the quality of life in each neighbourhood. There are 140 neighbourhoods (observations) per dataset and different features. The main dataset contains necessary information like Neighbourhood Id, Borough, latitude, and longitude. Although all 140 neighbourhoods are processed for each dataset, only 40 neighbourhoods are selected for this assessment. The 40 neighbourhoods selected are based on Toronto’s location. The neighbourhoods are closest to the center of the Canadian city.

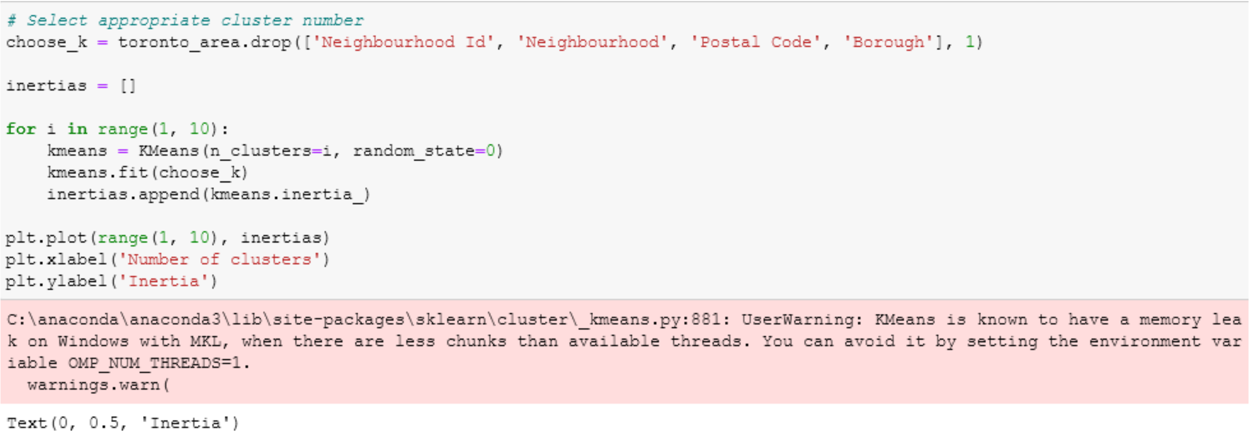


```
New Toronto dataframe shape: (40, 6)
<class 'pandas.core.frame.DataFrame'>
```

	Cluster Labels	Neighbourhood Id	Neighbourhood	Borough	Postal Code	Latitude	Longitude
62	1	63	The Beaches	East Toronto	M4E	43.676357	-79.293031
63	1	64	Woodbine Corridor	East Toronto	M4E	43.676357	-79.293031
64	1	65	Greenwood-Coxwell	East Toronto	M4L	43.668999	-79.315572
65	1	66	Danforth	East Toronto	M4C	43.695344	-79.318389
66	4	67	Playter Estates-Danforth	East Toronto	M4K	43.679557	-79.352188
67	4	68	North Riverdale	East Toronto	M4K	43.679557	-79.352188
68	4	69	Blake-Jones	East Toronto	M4J	43.685347	-79.338106
69	4	70	South Riverdale	East Toronto	M4K	43.679557	-79.352188

KMeans Clustering

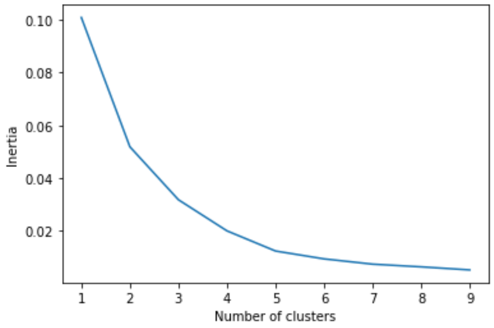
An unsupervised machine learning algorithm is utilized in this assessment to create the necessary cluster labels, which are clustered based on the latitude and longitude data. In order to determine the amount of optimal clusters, the elbow method is calculated. The lowest point or optimal point is where the inertia is closest to the center of the cluster – 5.



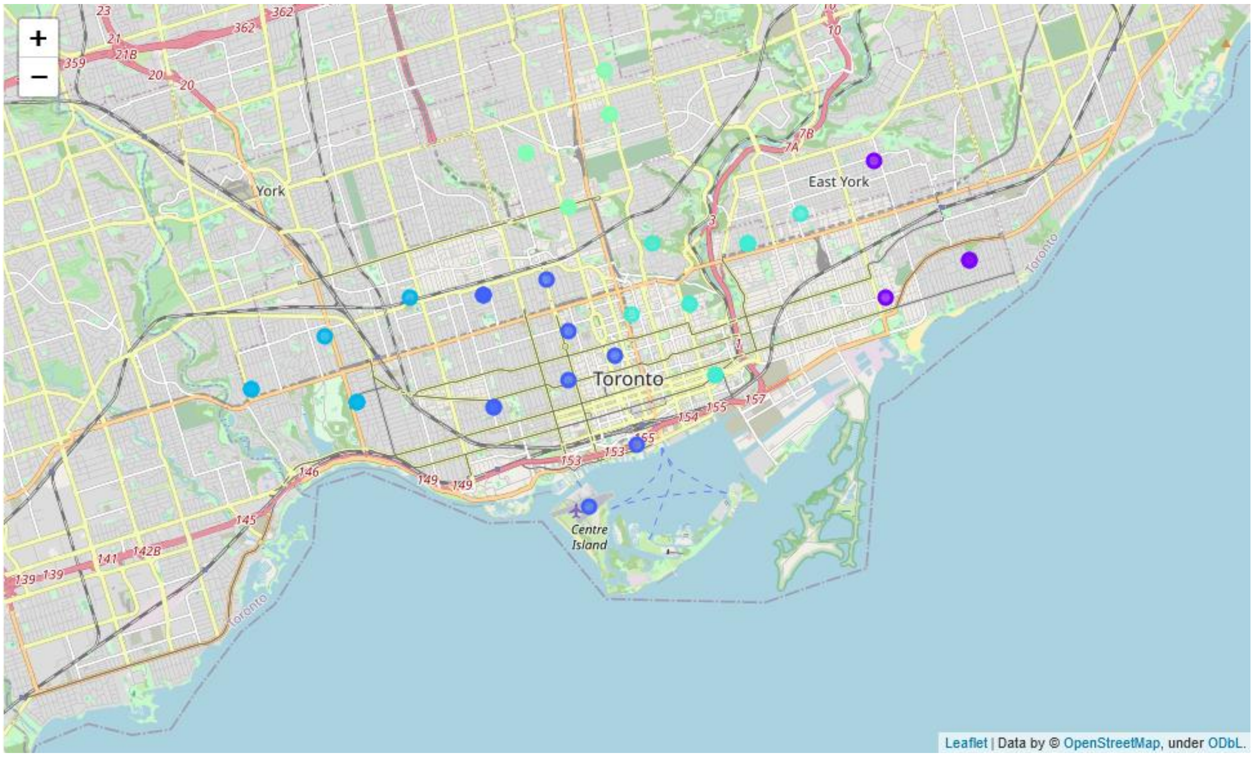
```
# Select appropriate cluster number
choose_k = toronto_area.drop(['Neighbourhood Id', 'Neighbourhood', 'Postal Code', 'Borough'], 1)

inertias = []

for i in range(1, 10):
    kmeans = KMeans(n_clusters=i, random_state=0)
    kmeans.fit(choose_k)
    inertias.append(kmeans.inertia_)

plt.plot(range(1, 10), inertias)
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
```

```
C:\anaconda\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:881: UserWarning: KMeans is known to have a memory lea
k on Windows with MKL, when there are less chunks than available threads. You can avoid it by setting the environment var
iable OMP_NUM_THREADS=1.
  warnings.warn(
```

```
Text(0, 0.5, 'Inertia')
```



Visually, each chosen neighbourhood is clustered in 5 cluster labels (between 0-4). East Toronto, Downtown Toronto, West Toronto, and Central Toronto are the boroughs selected.



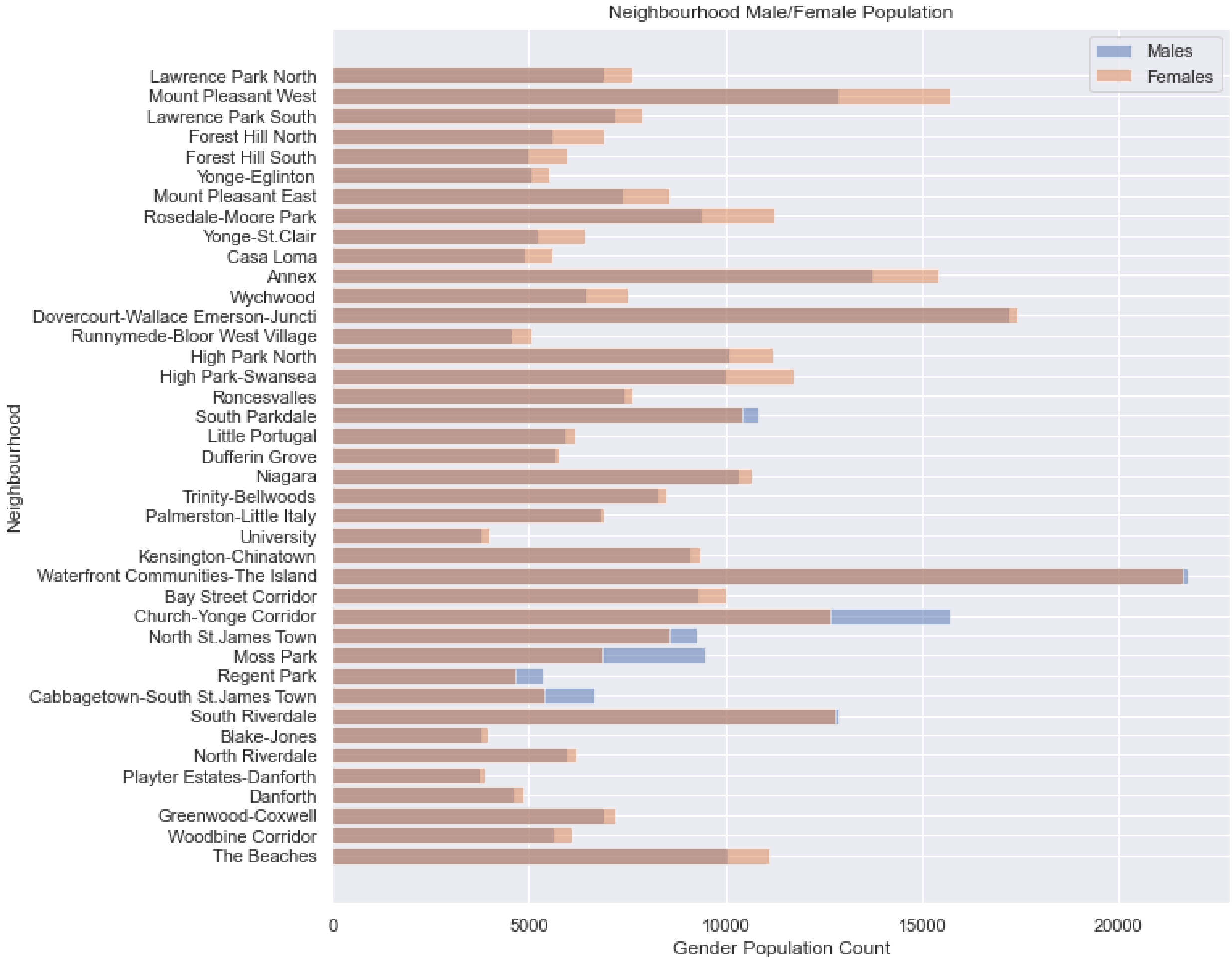
Data Notice

During data cleaning, all datasets have been pre-explored and may contain zero values that are included. These zero values may have not been collected by Toronto's Open Data Platform. This indicates that some features may not be represented or accurate during data collection. The values remain in this assessment, since other features in each datasets may be useful and create a more accurate picture during data exploration. Zero values will not be part of the final selection of the neighbourhood. The majority of the datasets were collected in 2011. These datasets may not be entirely accurate in comparison to today's data. Newer data will be added towards the end of the neighbourhood selection process.

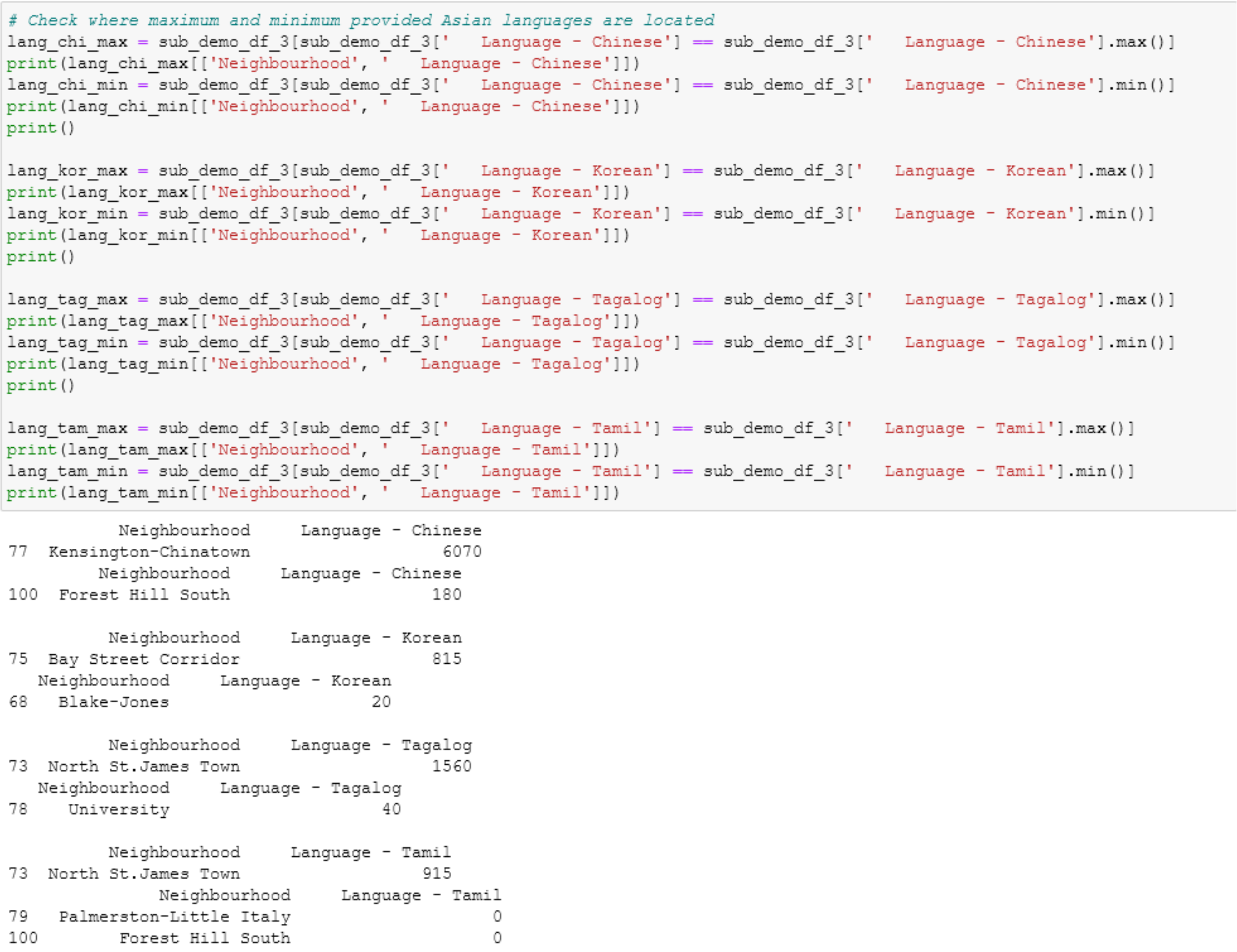
Data Exploration 1.0

The standard Maximum, Mean, Minimum, and visualization techniques were utilized in this assessment. Such techniques are used to help benefit each stakeholder as they may not be too technical to understand other data science techniques. All visual graphs uses the horizontal bar plot for ease of use and consistency in the assessment.

The chosen neighbourhoods are not as crowded as other cities – less than 25000 in each neighbourhood. The dataset also shows that the majority of the neighbourhoods have more females than males.



Not all Asian languages are recorded in the datasets. It seems only to include Chinese, Korean, Tagalog, and Tamil languages are the most prominent languages in Toronto in 2011.

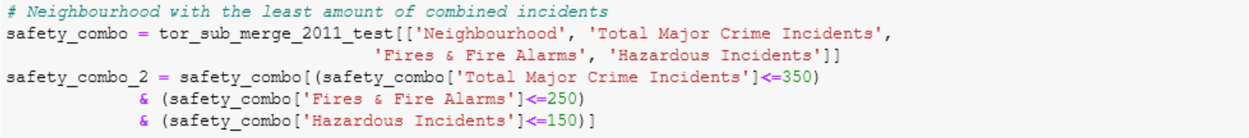


```
# Check where maximum and minimum provided Asian languages are located
lang_chi_max = sub_demo_df_3[sub_demo_df_3['   Language - Chinese'] == sub_demo_df_3['   Language - Chinese'].max()]
print(lang_chi_max[['Neighbourhood', '   Language - Chinese']])
lang_chi_min = sub_demo_df_3[sub_demo_df_3['   Language - Chinese'] == sub_demo_df_3['   Language - Chinese'].min()]
print(lang_chi_min[['Neighbourhood', '   Language - Chinese']])
print()

lang_kor_max = sub_demo_df_3[sub_demo_df_3['   Language - Korean'] == sub_demo_df_3['   Language - Korean'].max()]
print(lang_kor_max[['Neighbourhood', '   Language - Korean']])
lang_kor_min = sub_demo_df_3[sub_demo_df_3['   Language - Korean'] == sub_demo_df_3['   Language - Korean'].min()]
print(lang_kor_min[['Neighbourhood', '   Language - Korean']])
print()

lang_tag_max = sub_demo_df_3[sub_demo_df_3['   Language - Tagalog'] == sub_demo_df_3['   Language - Tagalog'].max()]
print(lang_tag_max[['Neighbourhood', '   Language - Tagalog']])
lang_tag_min = sub_demo_df_3[sub_demo_df_3['   Language - Tagalog'] == sub_demo_df_3['   Language - Tagalog'].min()]
print(lang_tag_min[['Neighbourhood', '   Language - Tagalog']])
print()

lang_tam_max = sub_demo_df_3[sub_demo_df_3['   Language - Tamil'] == sub_demo_df_3['   Language - Tamil'].max()]
print(lang_tam_max[['Neighbourhood', '   Language - Tamil']])
lang_tam_min = sub_demo_df_3[sub_demo_df_3['   Language - Tamil'] == sub_demo_df_3['   Language - Tamil'].min()]
print(lang_tam_min[['Neighbourhood', '   Language - Tamil']])
```

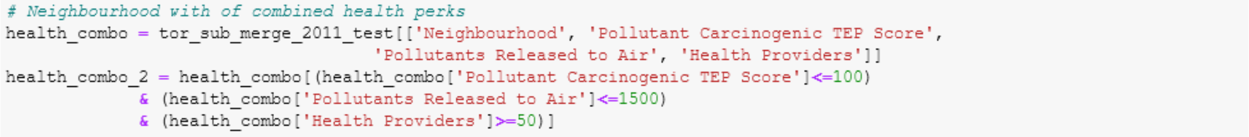
```
           Neighbourhood     Language - Chinese
77  Kensington-Chinatown                   6070
         Neighbourhood     Language - Chinese
100  Forest Hill South                    180

          Neighbourhood     Language - Korean
75  Bay Street Corridor                   815
   Neighbourhood     Language - Korean
68   Blake-Jones                    20

          Neighbourhood     Language - Tagalog
73  North St.James Town                   1560
   Neighbourhood     Language - Tagalog
78    University                     40

          Neighbourhood     Language - Tamil
73  North St.James Town                  915
               Neighbourhood     Language - Tamil
79   Palmerston-Little Italy                    0
100        Forest Hill South                    0
```

Data Exploration 2.0

The safety and health factor parameters are subset for further investigation. These factors are based on the stakeholder.

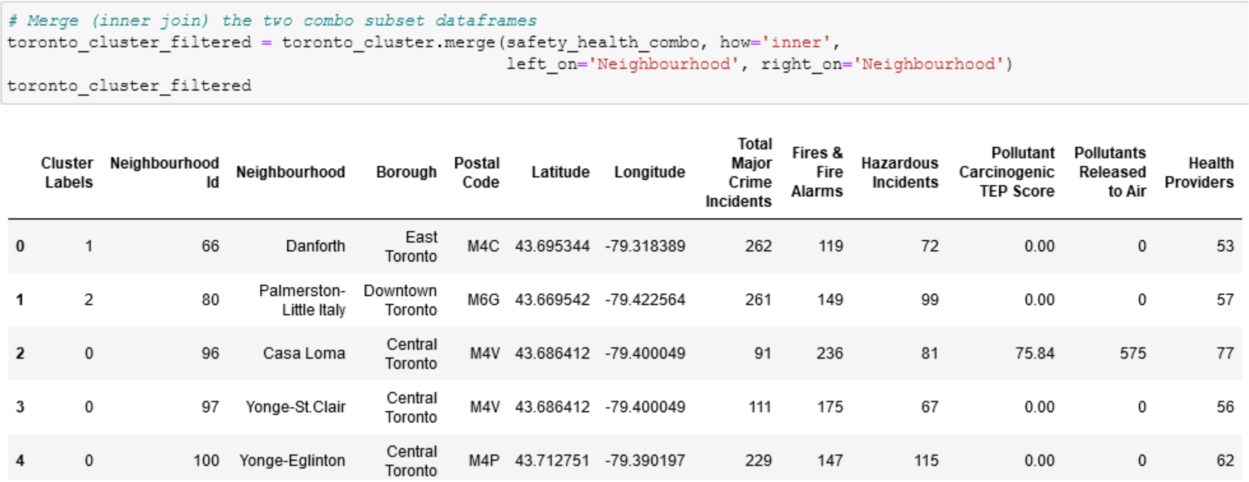


```
# Neighbourhood with the least amount of combined incidents
safety_combo = tor_sub_merge_2011_test[['Neighbourhood', 'Total Major Crime Incidents',
                                     'Fires & Fire Alarms', 'Hazardous Incidents']]
safety_combo_2 = safety_combo[(safety_combo['Total Major Crime Incidents']<=350)
             & (safety_combo['Fires & Fire Alarms']<=250)
             & (safety_combo['Hazardous Incidents']<=150)]
```



```
# Neighbourhood with of combined health perks
health_combo = tor_sub_merge_2011_test[['Neighbourhood', 'Pollutant Carcinogenic TEP Score',
                                     'Pollutants Released to Air', 'Health Providers']]
health_combo_2 = health_combo[(health_combo['Pollutant Carcinogenic TEP Score']<=100)
             & (health_combo['Pollutants Released to Air']<=1500)
             & (health_combo['Health Providers']>=50)]
```

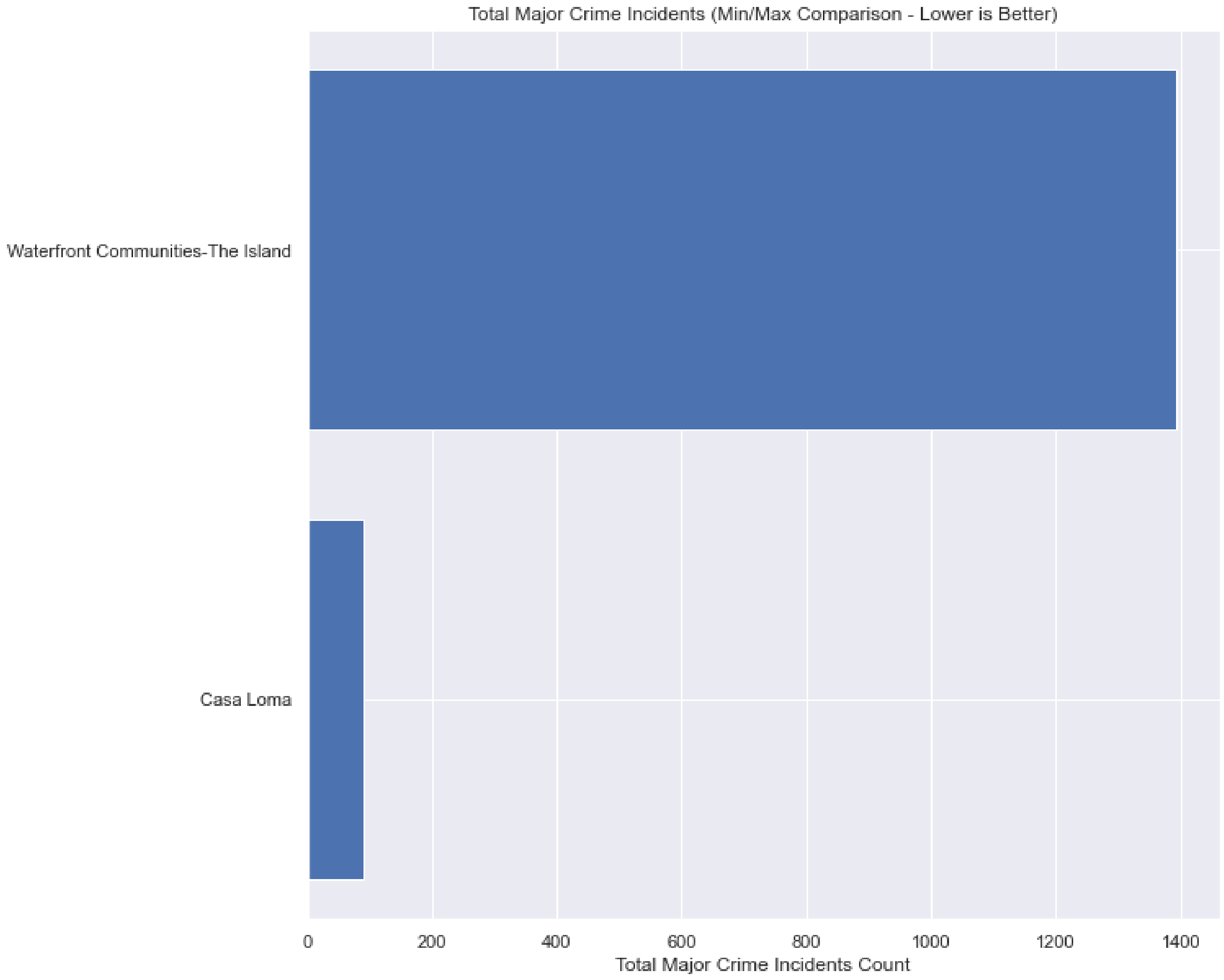
Both DataFrames are then combined. As mentioned in the Data Notice section, the zero values are going to be avoided in the neighbourhood process. The zero values are kept, since other features provide non-zero values that are part of that feature. Hence, the only neighbourhood available as a choice is the neighbourhood of Casa Loma.

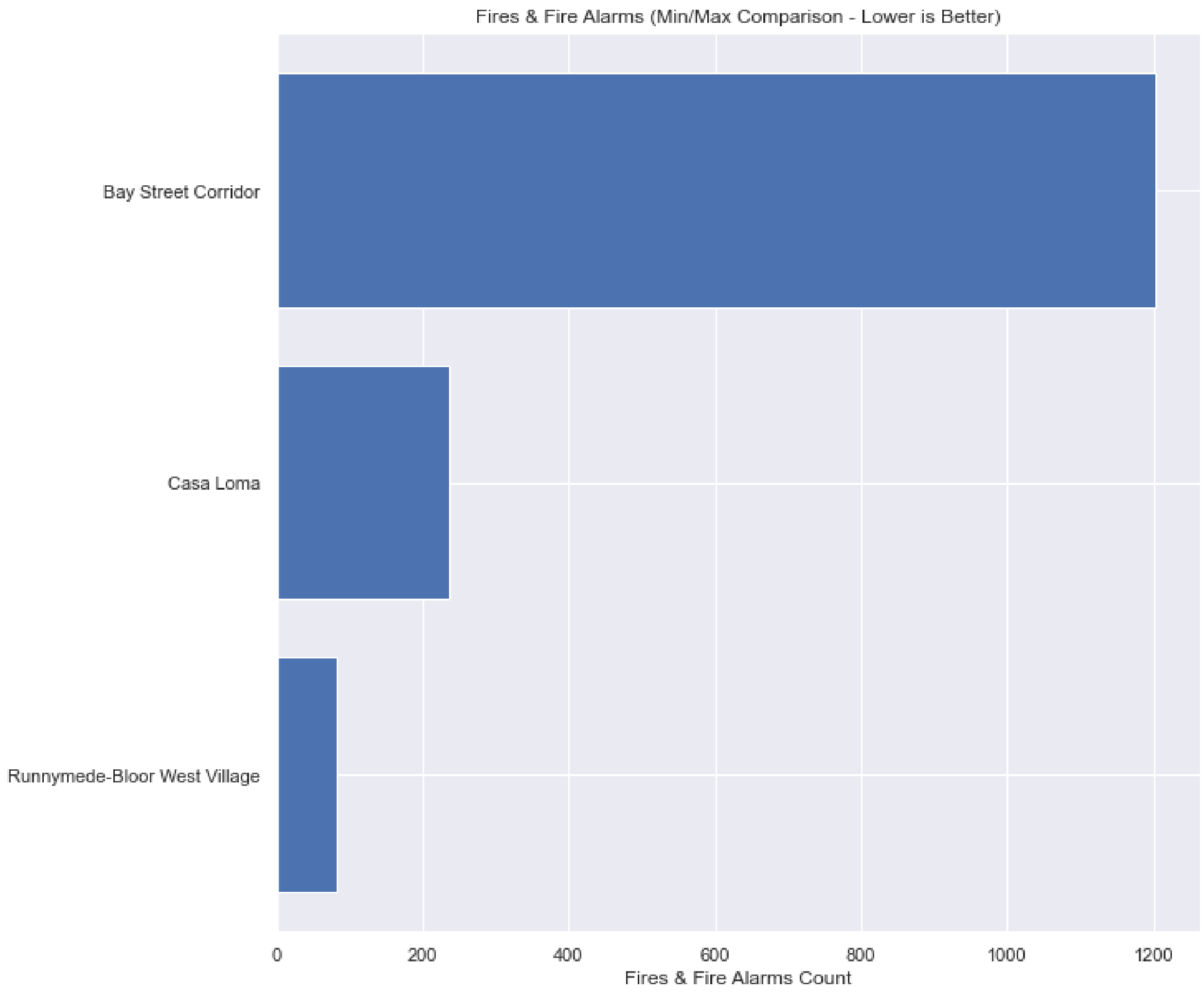


```
# Merge (inner join) the two combo subset dataframes
toronto_cluster_filtered = toronto_cluster.merge(safety_health_combo, how='inner',
                                                 left_on='Neighbourhood', right_on='Neighbourhood')
toronto_cluster_filtered
```

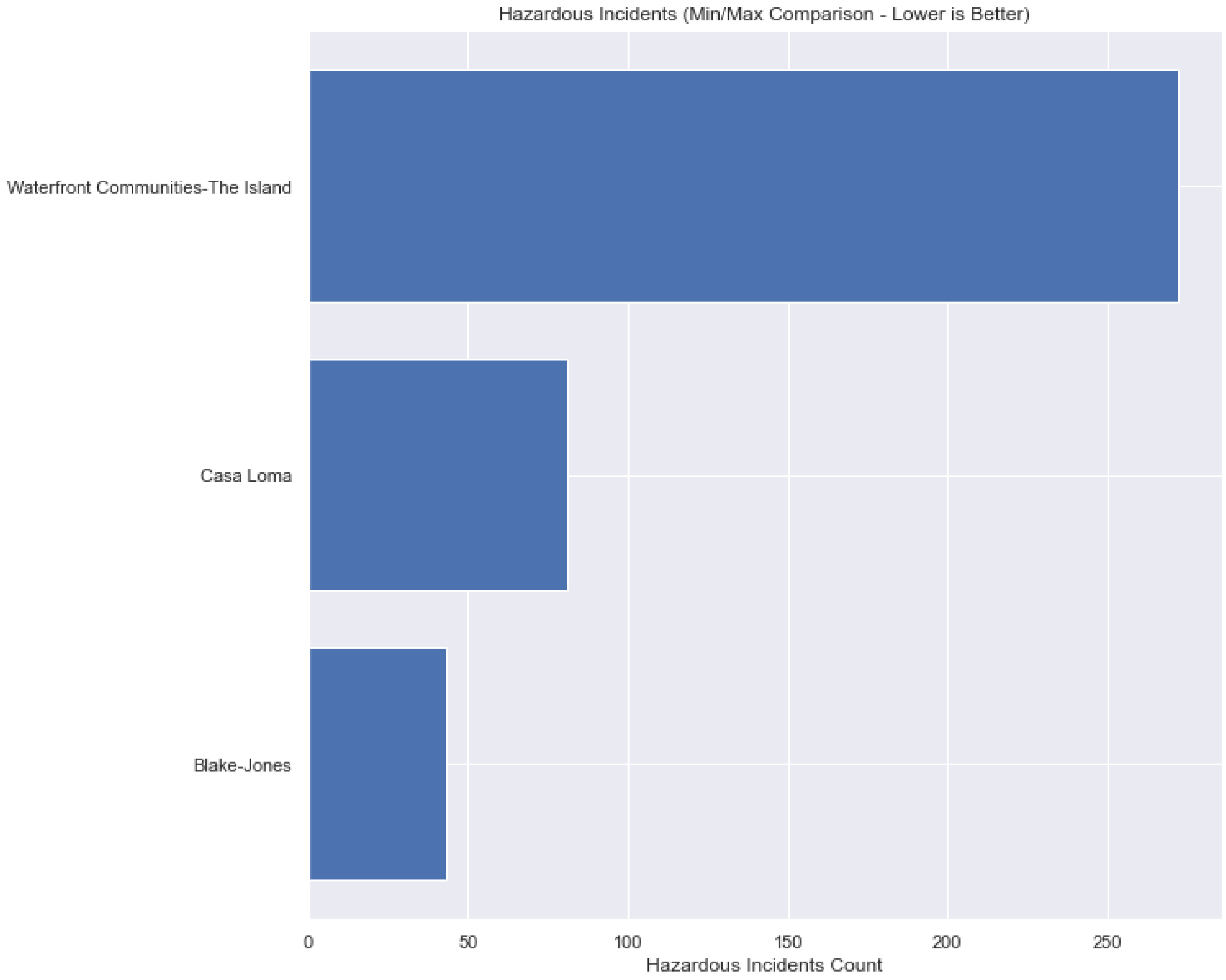
	Cluster Labels	Neighbourhood Id	Neighbourhood	Borough	Postal Code	Latitude	Longitude	Total Major Crime Incidents	Fires & Fire Alarms	Hazardous Incidents	Pollutant Carcinogenic TEP Score	Pollutants Released to Air	Health Providers
0	1	66	Danforth	East Toronto	M4C	43.695344	-79.318389	262	119	72	0.00	0	53
1	2	80	Palmerston-Little Italy	Downtown Toronto	M6G	43.669542	-79.422564	261	149	99	0.00	0	57
2	0	96	Casa Loma	Central Toronto	M4V	43.686412	-79.400049	91	236	81	75.84	575	77
3	0	97	Yonge-St.Clair	Central Toronto	M4V	43.686412	-79.400049	111	175	67	0.00	0	56
4	0	100	Yonge-Eglinton	Central Toronto	M4P	43.712751	-79.390197	229	147	115	0.00	0	62

Verifying the assessment with Maximum, Mean, Minimum techniques per health and safety factor.

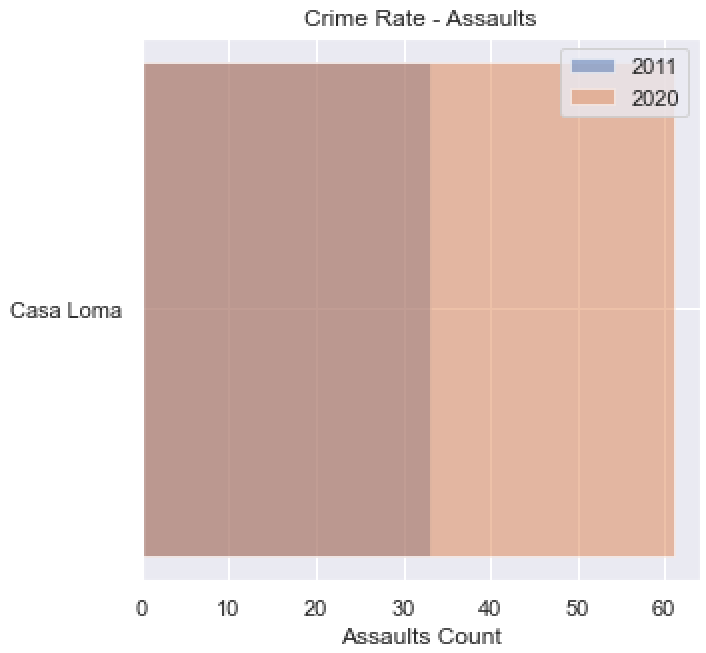




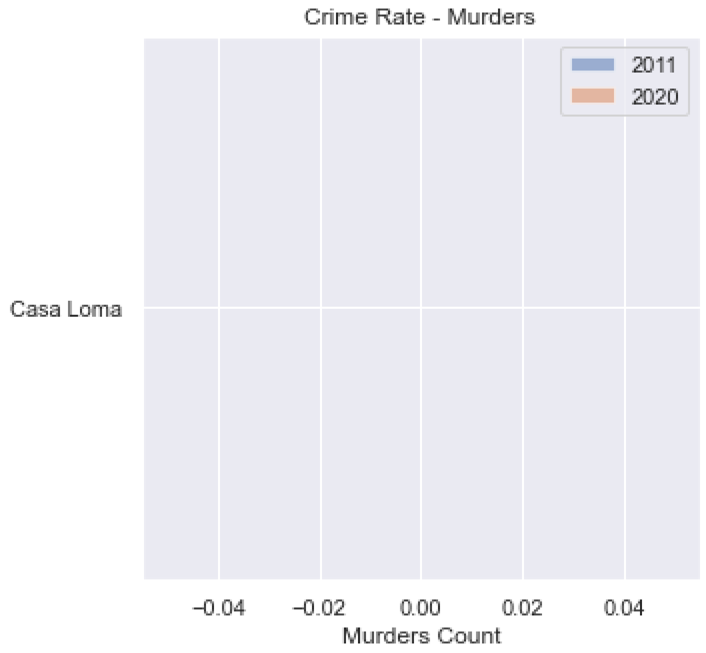
NOTE: Not all safety and health factors will be added to this report. To view more information about other factor measurements, please view the Jupyter notebook included in this project assessment.



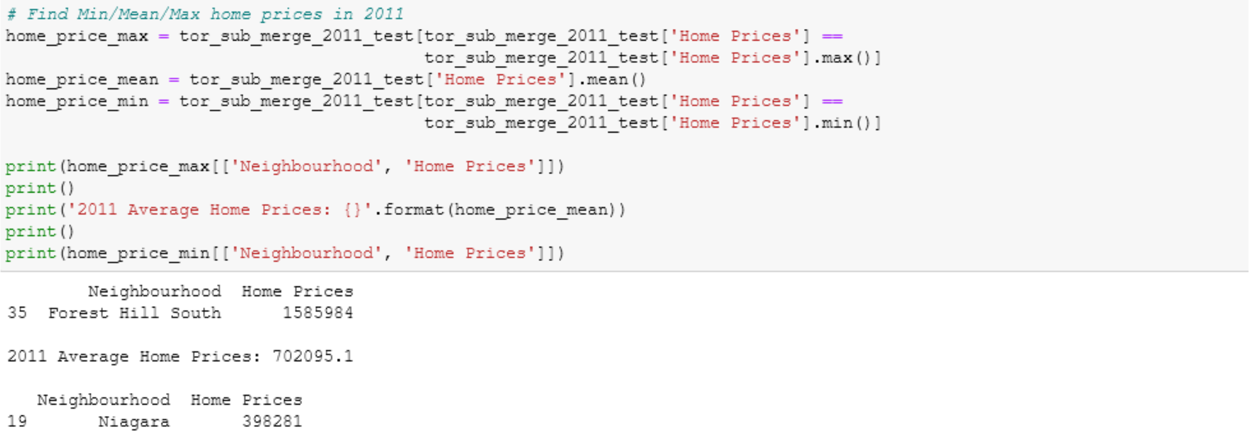
Updated data is available for the safety factors, which should also provide a more up-to-date measurements the stakeholder can review. The following horizontal bar plots assesses each major crime in the Casa Loma neighbourhood. The horizontal bar plots include measurement on assaults, vehicle thefts, breaking and entering, robberies, thefts, and murders. The murder horizontal bar plot is empty, since there was no change and no murder was committed in the neighbourhood.



NOTE: Not all the 2011 through 2020 crime rate data will be added to this report. To view more information about other year-to-year crime rate measurements, please view the Jupyter notebook included in this project assessment.

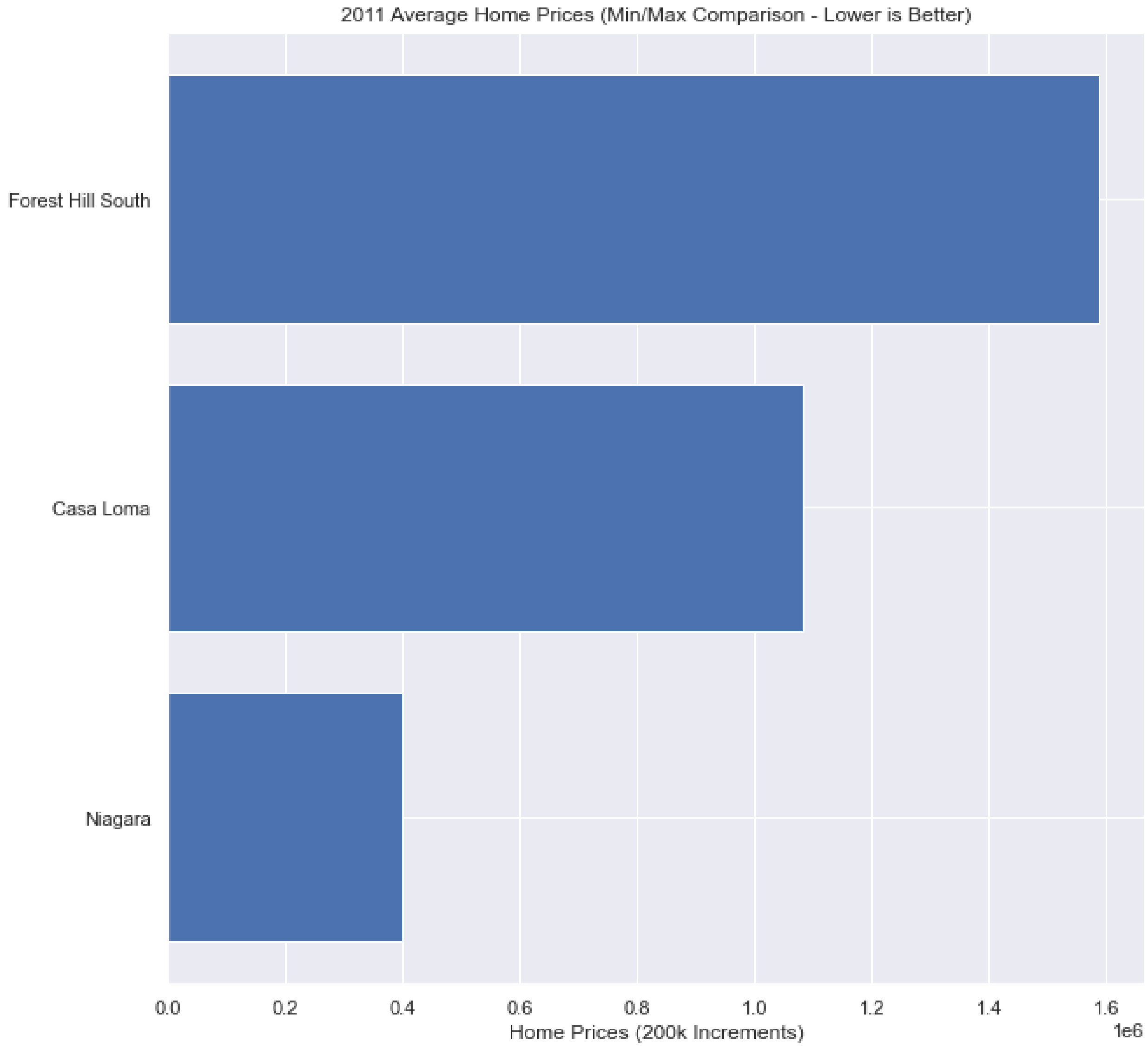


Outside of the safety and health factors defined in the Background section, home price may be another factor that can be a concern to any stakeholder/s. As of 2021, the Canada's inflation rate is at 3.4%. Considering the current price from 2011 in today's market is important. The Casa Loma neighbourhood's 2011 average home price is above average. Also note that, the prices are in Canadian currency prices.



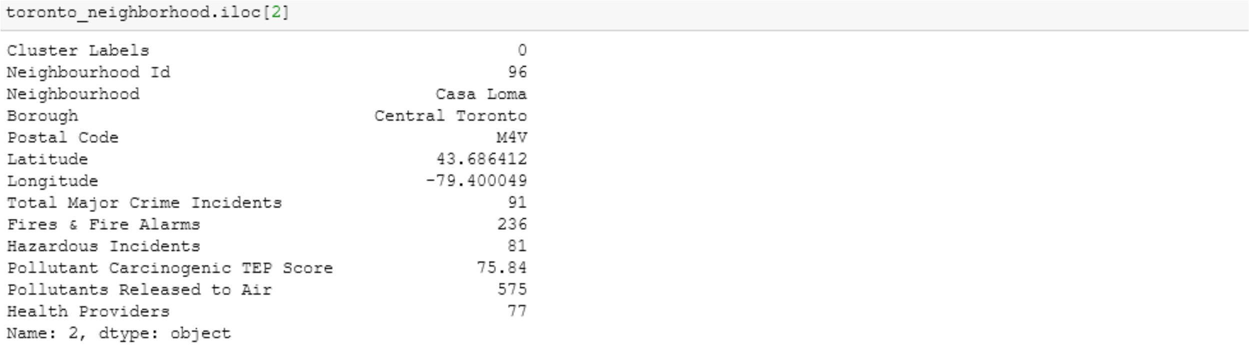
```python
# Find Min/Mean/Max home prices in 2011
home_price_max = tor_sub_merge_2011_test[tor_sub_merge_2011_test['Home Prices'] ==
                                          tor_sub_merge_2011_test['Home Prices'].max()]
home_price_mean = tor_sub_merge_2011_test['Home Prices'].mean()
home_price_min = tor_sub_merge_2011_test[tor_sub_merge_2011_test['Home Prices'] ==
                                          tor_sub_merge_2011_test['Home Prices'].min()]

print(home_price_max[['Neighbourhood', 'Home Prices']])
print()
print('2011 Average Home Prices: {}'.format(home_price_mean))
print()
print(home_price_min[['Neighbourhood', 'Home Prices']])
```

```
        Neighbourhood  Home Prices
35  Forest Hill South      1585984

2011 Average Home Prices: 702095.1

   Neighbourhood  Home Prices
19       Niagara       398281
```



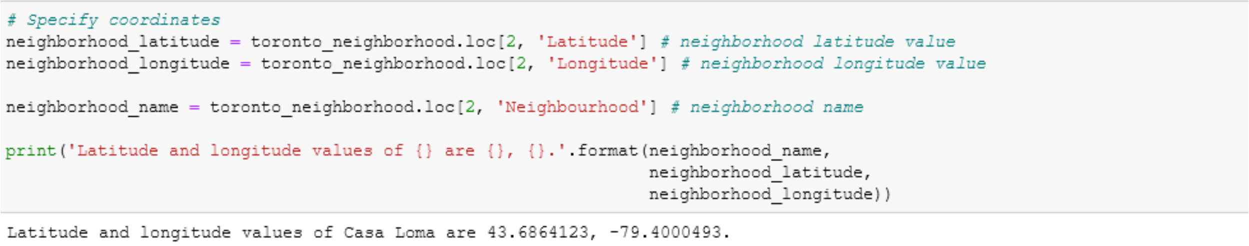
Foursquare Data Assessment

The Foursquare dataset will be utilized to determine where restaurants are located and what type of restaurants may be competitive in that neighbourhood. Specifying the neighbourhood latitude and longitude location is next, then followed by creating a search query for Thai restaurants using the Foursquare API.



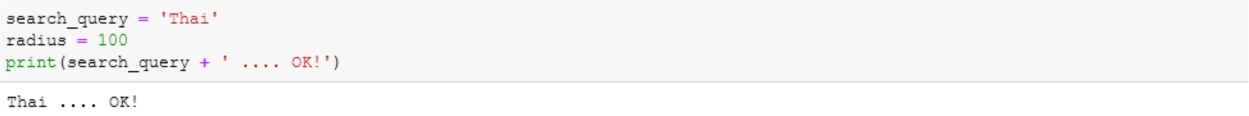
```
toronto_neighborhood.iloc[2]
```

```
Cluster Labels                                        0
Neighbourhood Id                                     96
Neighbourhood                                 Casa Loma
Borough                                 Central Toronto
Postal Code                                         M4V
Latitude                                      43.686412
Longitude                                    -79.400049
Total Major Crime Incidents                          91
Fires & Fire Alarms                                 236
Hazardous Incidents                                  81
Pollutant Carcinogenic TEP Score                  75.84
Pollutants Released to Air                          575
Health Providers                                     77
Name: 2, dtype: object
```



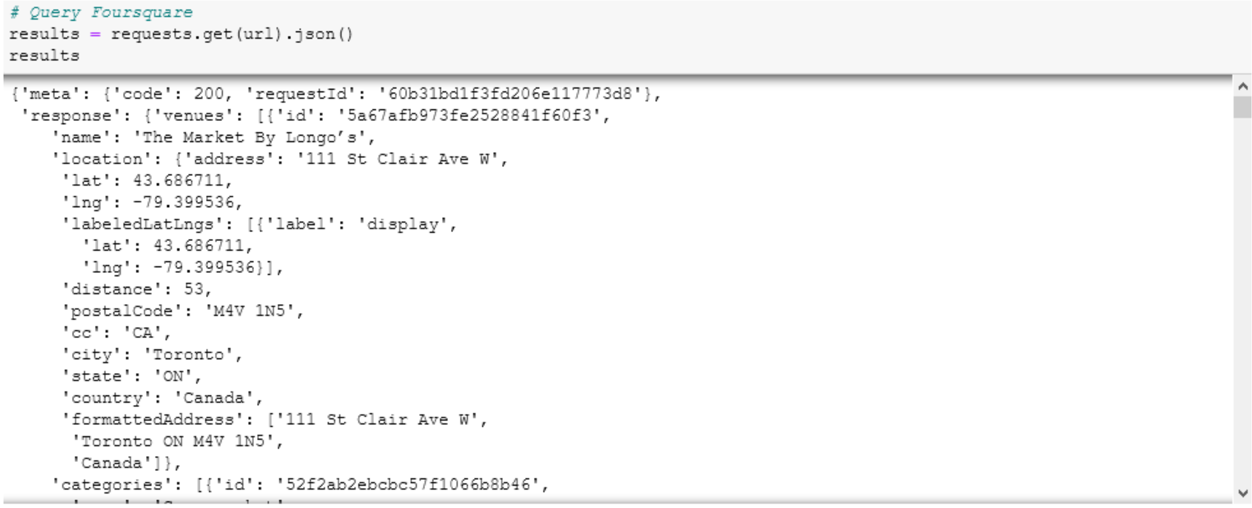
```
# Specify coordinates
neighborhood_latitude = toronto_neighborhood.loc[2, 'Latitude'] # neighborhood latitude value
neighborhood_longitude = toronto_neighborhood.loc[2, 'Longitude'] # neighborhood longitude value

neighborhood_name = toronto_neighborhood.loc[2, 'Neighbourhood'] # neighborhood name

print('Latitude and longitude values of {} are {}, {}.'.format(neighborhood_name,
                                                               neighborhood_latitude,
                                                               neighborhood_longitude))
```

```
Latitude and longitude values of Casa Loma are 43.6864123, -79.4000493.
```



```
search_query = 'Thai'
radius = 100
print(search_query + ' .... OK!')
```

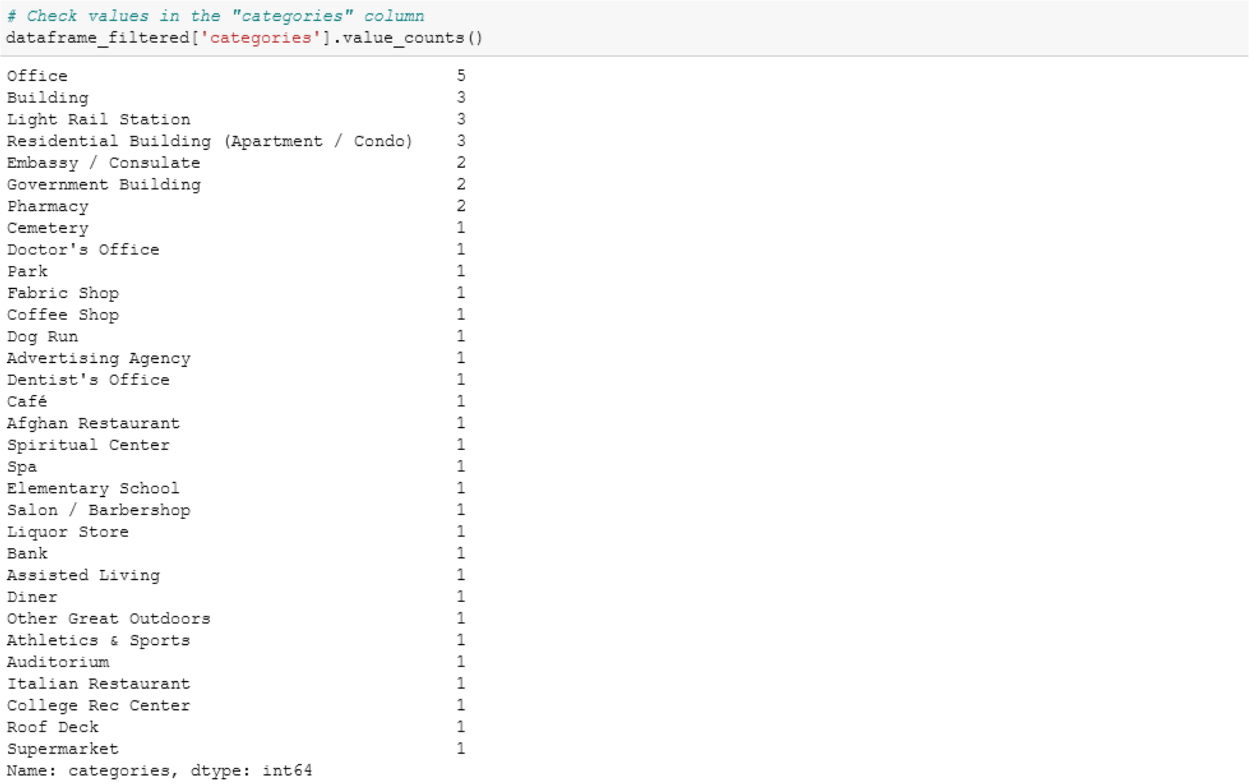
```
Thai .... OK!
```



```
# Query Foursquare
results = requests.get(url).json()
results
```

```
{'meta': {'code': 200, 'requestId': '60b31bd1f3fd206e117773d8'},
 'response': {'venues': [{'id': '5a67afb973fe2528841f60f3',
    'name': 'The Market By Longo's',
    'location': {'address': '111 St Clair Ave W',
     'lat': 43.686711,
     'lng': -79.399536,
     'labeledLatLngs': [{'label': 'display',
       'lat': 43.686711,
       'lng': -79.399536}],
     'distance': 53,
     'postalCode': 'M4V 1N5',
     'cc': 'CA',
     'city': 'Toronto',
     'state': 'ON',
     'country': 'Canada',
     'formattedAddress': ['111 St Clair Ave W',
      'Toronto ON M4V 1N5',
      'Canada']},
    'categories': [{'id': '52f2ab2ebcbc57f1066b8b46',
```

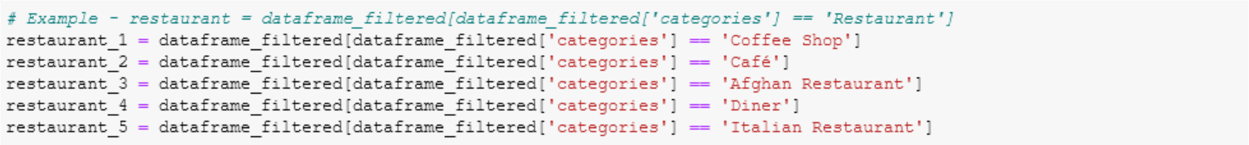
Filtering the JSON output and transforming it to a DataFrame is next. Filtering the categories column in the DataFrame should provide what food venues are available within a 100-mile radius.



```
# Check values in the "categories" column
dataframe_filtered['categories'].value_counts()
```

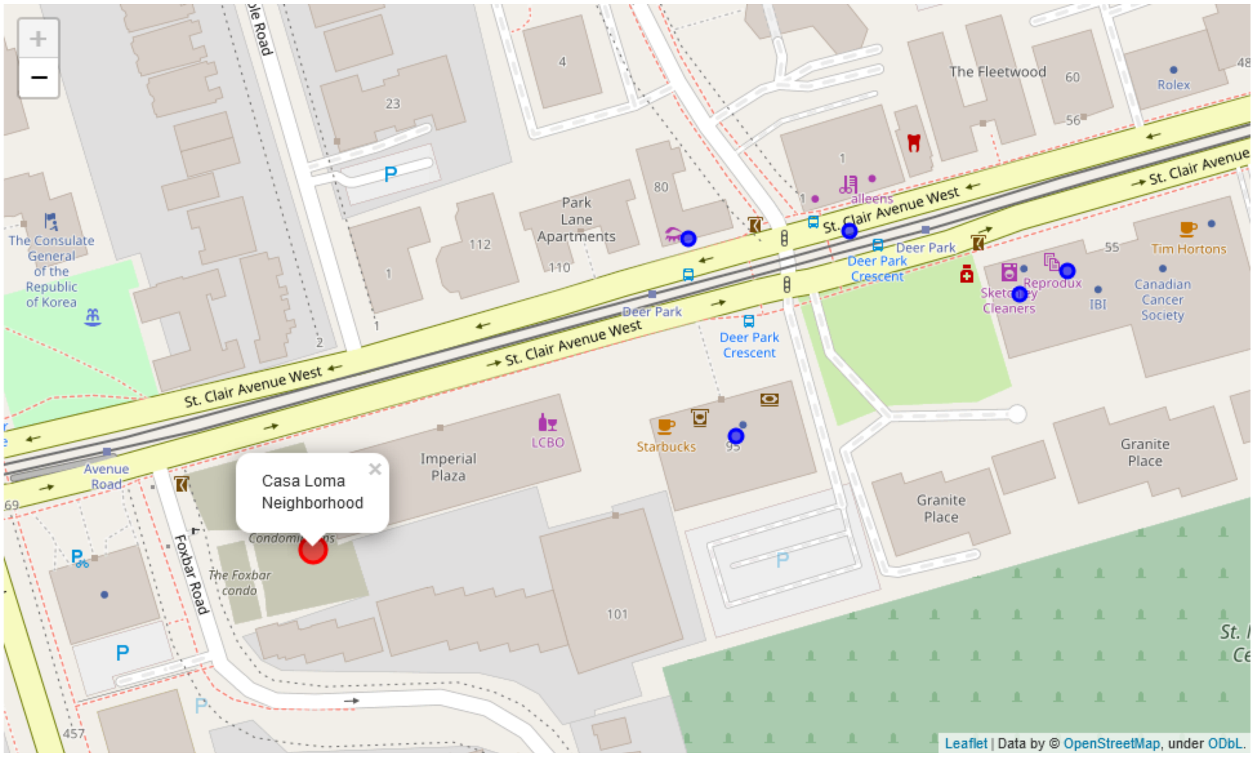
```
Office                                     5
Building                                   3
Light Rail Station                         3
Residential Building (Apartment / Condo)   3
Embassy / Consulate                        2
Government Building                        2
Pharmacy                                   2
Cemetery                                   1
Doctor's Office                            1
Park                                       1
Fabric Shop                                1
Coffee Shop                                1
Dog Run                                    1
Advertising Agency                         1
Dentist's Office                           1
Café                                       1
Afghan Restaurant                          1
Spiritual Center                           1
Spa                                        1
Elementary School                          1
Salon / Barbershop                         1
Liquor Store                               1
Bank                                       1
Assisted Living                            1
Diner                                      1
Other Great Outdoors                       1
Athletics & Sports                         1
Auditorium                                 1
Italian Restaurant                         1
College Rec Center                         1
Roof Deck                                  1
Supermarket                                1
Name: categories, dtype: int64
```

The top 50 search query on the Foursquare API did not produce Thai-based food venue in the neighbourhood. The result is a good sign that this neighbourhood is indeed a viable choice to open a Thai restaurant.



```
# Example - restaurant = dataframe_filtered[dataframe_filtered['categories'] == 'Restaurant']
restaurant_1 = dataframe_filtered[dataframe_filtered['categories'] == 'Coffee Shop']
restaurant_2 = dataframe_filtered[dataframe_filtered['categories'] == 'Café']
restaurant_3 = dataframe_filtered[dataframe_filtered['categories'] == 'Afghan Restaurant']
restaurant_4 = dataframe_filtered[dataframe_filtered['categories'] == 'Diner']
restaurant_5 = dataframe_filtered[dataframe_filtered['categories'] == 'Italian Restaurant']
```

The map also suggests that each food venue found are clustered or fairly near each other. This may indicate that the area may be a hotspot for everyone in the neighbourhood to shop and eat. Possibly placing the Thai restaurant nearby the other food venues may be a good idea.



Results and Discussion

The analysis provided favorable output for the stakeholder/s, where there is one neighbourhood that meets their primary criteria of safety and health factors. The Foursquare API results also provided a favorable outcome as the search query has shown no competition in the same ethnic restaurant theme/choice - Thai cuisine. Not only was the Toronto datasets and Foursquare outcome presented the best neighbourhood in Toronto, those results has also provided the stakeholder/s with viable information that make up that neighbourhood. For example, there is a low density of restaurants found within a 100-mile radius. Meaning a higher chance of successfully launching the Thai restaurant. The neighbourhood of Casa Loma also have less crime-rated offenses compared to other neighbourhoods. And less pollution produced which have a higher chance of remaining healthier compare to other neighbourhoods.

One caveat with the analysis is the use of available datasets. The datasets utilized are nearly from a decade ago, 2011. Approximately 10 years have passed and Toronto may have changed drastically over the years (e.g. more population, more neighbourhood division - 158 from 140, et al).Not to mention, that the dataset can be incomplete with missing data, which can potentially limit the choices in the data exploration and analysis process. Another caveat for the stakeholder/s is the home price cost in Casa Loma. The cost is above average and that type of price can be one of the main factors that can shape this type of assessment for any stakeholder considering that the inflation rate has increased in the last 10 years - 3.4%.

Conclusion

The purpose of the assessment is to identify potential neigbourhoods to start a Thai restaurant business at mainly based on health and safety factors. The assessment was able to determine a set of favorable neighbourhoods with tolerable health and safety factors and one neighbourhood stood out. The neighbourhood of Casa Loma is recommended as this neighbourhood provides the stakeholder/s with a complete picture of how the neighbourhood is like, which provides them an ideal neighbourhood to start a business in and to live nearby. Other factors such as home price, school attractiveness, local attractions, and etc. may also be factors, but will ultimately be decided by stakeholder/s. For future tasks, the Open Data Platform has announced that new data is incoming. Updating the data gathered for each neighbourhood can provide more investigation via data comparison or even data correlation as needed.